Stats 2: Study Designs and Measures: USMLE Step 1 Review Material

Objectives:

1. Know the concepts and calculations covered on the USMLE Step 1
2. Be able to define epidemiology and describe the basic epidemiology study designs
3. Calculate and interpret the following: risk and relative risk (RR) for a cohort study, and an odds ratio (OR) for a case-control study

Epidemiology

* The study of how disease is distributed in populations and the factors that influence or determine this distribution
* The study of the distribution and determinants of health-related states or events in specified populations and the application of this study to control of health problems
* Questions to ask include: Why does a disease develop in some people and not in others? The premise of epidemiology is that disease, illness, and ill health are not randomly distributed in a population. We all have certain characteristics that predispose or protect us from a variety of different diseases.
* Definition number two covers the content and purpose of epidemiology.

What are we trying to accomplish?

1. Identify risk factors of diseases
   * Risk factors increase a person’s likelihood of acquiring a disease, condition, or event. We are trying to identify causes, though this really is not possible. What we really try to determine are RISK FACTORS. The goals being to know how a disease/condition/event develops and how it might be transmitted (more related to infectious diseases).
   * The goals of epidemiology are to reduce morbidity and mortality and to develop prevention and intervention programs
2. Determine the extent of disease found in a community
   * We want to know the burden (person, place, time) of disease as this helps for planning health services, facilities, and training of health care providers. Surveillance of populations (monitoring statistics of disease/events) is useful for this.
3. Study natural history and prognosis of disease
   * so we can create modes of intervention, including treatment and prevention of complications
4. Evaluate new preventive and therapeutic measures and new modes of health delivery
5. Provide foundation for developing public policy and regulatory decisions
   * May deal with environmental problems such as radiation, radon in homes, occupations associated with an increased risk of disease in workers. How could we regulate conditions in the workplace? This is where policies are needed.

Goals of epidemiology:

1. Identify subgroups in the population who are at high risk for disease

* If we can identify these groups, we may be able to identify the risk factors that increase their risk of disease and thus modify exposure to those risk factors.
* Direct preventive efforts, such as screening for early detection) to populations are useful to those who can benefit from intervention.

1. Primary, secondary, and tertiary prevention

* Primary prevention is preventing development of a disease in a person who is well. Again, this does not have to be a disease, it can be an event such as preventing injuries in the workplace (i.e., wearing hard hats in construction areas).

1. Reduce morbidity and mortality
2. Develop prevention and intervention programs
3. Improve patient prognosis by discovering methods to enhance quantity and quality of life

Randomized trials: aka intervention, analytic and experimental trials

* Causal Inference
  + Causation cannot be observed. Causation is **inferred**. Inferences depend upon conceptual models and frames of reference.
* The counterfactual Ideal
  + The counterfactual ideal means that the perfect comparison group for a group of exposed individuals consists of exactly the same individuals had they not been exposed. This would allow us to infer causation. Since it is not possible to compare the same people with and without the exposure at the same time and with everything else the same, this comparison is counterfactual. Only one condition can exist (i.e. being exposed) while the comparison condition (i.e. being unexposed) is counterfactual. Thus, intervention studies try to get as close as possible to this counterfactual condition through random allocation.
  + Simply put, what would have been the outcome in the exposed group if they had been unexposed?
  + Instead, we do the ‘best we can’ and compare the outcome for the exposed group to the outcome in a “substitute” population.
    - Substitute population represents the exposed group **without** the exposure
  + The idea is that the outcomes for the substitute population should be the same as the outcomes in the counterfactual comparison. In order to make this inference, we must find a valid substitute population.
  + This substitute population will come from your study population. The study population consists of the people who are considered for enrollment into the trial. It will therefore include those who get the treatment (exposure) and the substitute population (which in this example will not get the exposure). The study population depends on the purpose of the trial and its characteristics are defined by specific eligibility criteria.
  + Control cohort simply observe. For certain things eg can’t assign preg women to consume alcohol. So can’t do a controlled experiment.

Diff bw intervention trials and observational studies:

* Intervention (Trials)
  + Exposure status determined by researcher
  + Researcher offers an exposure to some subjects in order to observe some response
* Observational (Case-Control, Cohort, Cross-sectional, Ecological)
  + Exposure status ***not*** determined by researcher
  + Researcher observes subjects and measures variables of interest but does not impose any exposure
  + Can’t randomize certain things eg assign preg woman to drink alcohol. If they choose to do it then can study them

Intervention:

* The final proof in etiology – experimentation
* May involve removing or adding a factor and observing whether an outcome or a disease is prevented
* May use a surrogate outcome such as lower cholesterol, rather than prevent heart disease
* Some studies carried out to improve methods of treatment and prevention
* Others carried out for drug or vaccine licensure
* The best type of studies to study the etiology of health outcomes are of experimental nature, intervention (randomized) studies. The goal is to randomize everything possible except what you are trying to study, thus removing as much potential as possible for confounding. There is always the risk of unknown or unmeasured confounding (residual confounding) from a variable or variables that at the time are not known to be associated with both your exposure and outcome of interest.

Randomized Clinical Trials

* Trials of therapeutic interventions are ones in which

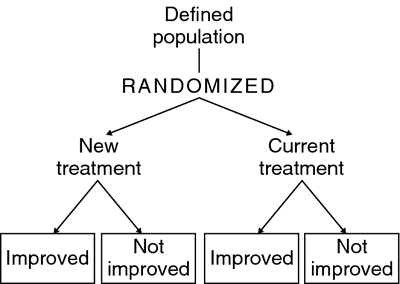
1. Patients are assigned to one of two or more groups to be offered different therapeutic measures
2. Chance alone dictates whether a particular patient will be assigned to a certain group
3. Patients in each group are monitored for the abatement of their illness, occurrence of the event(s) the therapy seeks to prevent, or occurrence of side effects
4. Although randomized controlled trials are important in identifying adverse effects of therapy that are relatively common and occur relatively soon after therapy has been initiated, the primary purpose of these trials is to assess the **efficacy of therapy**. Results of these trials can be interpreted much easier than those of non-randomized studies. This is because of the concern of non-randomized studies is that the various treatment groups had inherently unequal probabilities of doing well. In a **randomized** trial, it is only **chance** that determines the membership of the groups

Selecting intervention measure:

* Make the intervention to be studied as different as possible from the treatment received by the control patients
* It is easier for a randomized trial (or any study) to identify a large difference or association, rather than a small one (req smaller sample; small difference req larger sample)

Why randomize?

* To accurately estimate causal effects and make treatment groups comparable
* Best attempt to create a comparable substitute population (counterfactual)
* Why is comparability important?
  + The groups that we are comparing to each other could be different in some important way
  + What are common differences?
  + Propensity to volunteer
  + Tendency to comply
  + Healthy lifestyle
  + Genetic susceptibility
* Randomization is defined as an act of assigning or ordering that is the result of a random process. This is done after informed consent has been obtained. The two main goals of randomization are the following:
  1. To ensure that the treatment assignment occurs in an unbiased fashion
  2. If the sample size is sufficiently large, it will usually control for known and unknown confounding variables by producing treatment and comparison groups with very similar baseline characteristics (i.e. gender, race, age, SES).
* We randomize because it provides the strongest evidence for causal inference. This is because we are getting as close as possible to the counterfactual and the ideal situation of following the same people as exposed and unexposed at the same time.
* Randomization helps us control confounders that we cannot measure (potential examples, depending on the study are: genetic susceptibility, propensity to volunteer). Many outcomes are influenced by multiple factors. When we study the influence of one factor we obviously need to take into account, or control for, the influence of other factors. This will help isolate the relation of interest. With a sufficient sample size, randomization will help to ensure that potential confounders that are known to the investigators, as well as those unknown or even unsuspected by the investigators, are evenly distributed among the treatment groups. If there are imbalances in known or suspected risk factors due to small sample size or to the play of chance even in a large sample, these can be controlled in the data analysis by methods such as stratification and multivariate analysis.



How do we randomize:

* Simple random allocation
  + Roll of dice or random number table determines allocation to treatment or control for each subject
* Stratified random allocation
  + Subjects are first stratified by a baseline risk factor (e.g. smoking status)
  + Subjects in each stratum are randomized to treatment or control
* allocation into various treatment groups should take place only after subjects have been determined to be eligible and have expressed a willingness to participate. In other words, after informed consent has been obtained. To maximize the probability that the groups receiving the differing interventions will be comparable, assignment to a study group should be at random. Random assignment implies that each individual has the same chance of receiving each of the possible treatments and that the probability that a given subject will receive a particular allocation is independent of the probability that any other subject will receive the same treatment assignment. The most common methods used to achieve this objective are random number tables and computerized random number generators.
* Stratification helps ensure that key confounding variables are equally distributed between the treatment and comparison groups. Individuals are first stratified or separated according to the confounding characteristic (i.e., sex, age). Then, they are randomly assigned to either the treatment or comparison group within each stratum.
* Many times gender is stratified.

Blinding:

* Definition: participants (and investigators) are kept unaware whether they are in a treatment or control group
* Purpose: avoid actions that may influence measurement or interpretation as a result of awareness of the intervention; this bias can occur after study subjects have been randomized
* The likelihood of such bias is directly related to the subjectivity of the outcomes under study. If the endpoint being considered is total mortality, bias on the part of the investigator is unlikely since the fact of death is objective and indisputable. However, when trials include subjective outcomes such as severity of illness, frequency of side effects, decreased pain, and increased mobility, methods need to be in place to minimize the likelihood of any systematic difference in ascertaining outcomes between the study groups

Placebos and blinding

* Placebo effect: tendency to “experience” a favorable response if subjects think they are receiving a treatment
* Placebos: a sham treatment identical to real treatment but lacking active agent; equalizes placebo effect between groups
* Placebos prevent subjects from knowing whether they are in treatment or control group, facilitates blinding
* is intended to have no biologic effect outside of the other treatment itself. Placebo pills typically contain sugar or other essentially inert ingredients. Such pills can be manufactured to be distinguishable from the new therapy being offered.
* Although the use of a placebo facilitates blinding, this is not the primary use of a placebo. It has long been known that even if a treatment has no effect, offering that treatment may have a healthy effect. An offer of treatment is an offer of hope, and it may bring the expectation of treatment success. Expectations are thought to have a powerful influence on outcome. If so, a new treatment may have an effect that comes only through the lifting of patient expectations. The use of a placebo in a trial is intended to distinguish new treatments that have no more than the placebo effect from those that have more of an effect. The placebo effect itself is highly variable, depending on the nature of the outcome and the nature of the treatment.
* Hard to placebo a surgery

Randomized trials:

* Advantages
* It is an experiment
* Provides strongest evidence for causality in relation to temporality
* Reduces the influence of other determinants of exposure and outcome (confounding)
* Disadvantages
* Expensive
* Time consuming
* Subjects may not be representative of all people who might eventually be put on the treatment (Why?)
* Equipoise means there must be a genuine uncertainty about which treatment is better. That is the ethical basis for an intervention trial, that it is not possible to say whether the intervention is in fact better than the alternative.

Assessing baseline imbalances

* Baseline risk factors: risk and prognostic factors for the outcome that are measured at the time of random assignment
* If randomization succeeds in achieving its goal, the frequency of the outcome would be similar in the groups created by randomization, apart from the effect of the intervention
* Though there is no direct way to measure whether such a balance in overall prognosis for the groups has been achieved, it is possible to measure the distribution of individual prognostic factors in the compared groups to see how well balanced these are. Any imbalance in a baseline risk factor represents confounding, because it is associated with the exposure and the outcome.
* The outcome of a random process is predictable only if aggregated over many repetitions
* In a specific case or particular trial, unlikely distributions can result from the randomization
* Two distributions will rarely be identical, so how can we tell when the imbalance in a baseline risk factor is severe enough to warrant treating the variable as a confounding factor? If a factor that is severely imbalanced has only a small effect on the outcome, there will be little confounding even with the severe imbalance. On the other hand, even a minor imbalance in a strong risk factor for the outcome might lead to worrisome confounding. Thus, the amount of imbalance in the risk factor itself is not a good guide to the amount of confounding that the imbalance introduces. The way to assess confounding is to use the same approach that epidemiologists use in other situations, which is basically the method used to compare the crude estimate of effect with an adjusted estimate of effect (one that removes the confounding from that particular variable).

Mistakes in clinical trials: p-values

* Using statistical significance testing to assess imbalances in baseline risk factors is a common mistake.
* An imbalance in baseline risk factors by itself does not reflect the amount of confounding
* The amount of confounding is the result of the strength of the associations between the baseline risk factor and the exposure and outcome.
* However, the result of a statistical significance test depends not just on the strength of the association being tested but also on the size of the study. For a given strength of association, the more data, the smaller the p-value.
* Thus, a given amount of confounding in a large study might yield statistically significant differences, but the same amount of confounding in a small study may not.
* Instead, one should compare the crude effect estimate with the adjusted estimate (where confounding is controlled) and then assess the difference between the two results.
  + Another reason why the use of statistical significance tests to evaluate baseline imbalances in a clinical trial is improper is that you may want to know what the null hypothesis is. The answer must be that the null hypothesis is that any observed imbalance is just the result of chance. If a statistically significant result is observed, one would ordinarily take that to mean that the null hypothesis is rejected. In the case of baseline imbalances in a randomized trial, that would mean rejecting the hypothesis that chance produced the imbalances. But we cannot reject that hypothesis! Chance did in fact produce the imbalance: the imbalance is the result of a randomized allocation. Random assignment can produce unusual results, but we already know that in a trial the imbalances that do occur are due to chance. Therefore, it makes no sense to test the null hypothesis. What matters is that the imbalance exists and what is important to know is how much confounding it causes. Statistical significance testing cannot reveal that, but the straightforward application of epidemiologic rules for assessing confounding can.

Summary:

* Intervention trials are an experimental type of study design, and include clinical trials, field trials, and community intervention trials
* Randomization is a way of assigning study subjects to exposure groups that helps assure comparability between the groups
* The power a trial has to detect a difference between exposure groups depends on the sample size, acceptable error rates, and the magnitude of the difference you want to detect. Power will be discussed later in this class.
* Well-conducted intervention trials are thought to produce more scientifically rigorous results than do observational studies. This is because of the features of random assignment of participants to treatment and comparison groups to eliminate baseline differences between the groups and to reduce the likelihood of biased allocation, and the use of placebo controls to permit blinding for better ascertainment of outcomes.

Cohort studies (#2 to randomized)

* Cohort: any designated group of persons who are followed forward over a period of time. Fancy word for ‘group’
* Describes a group of people with a common characteristic or experience
* Cohort studies are considered observational studies because you are simply observing a population and collecting data
* Can calculate cumulative incidence (risk), relative risk (RR) and incidence rates (IR)
* Cohort studies of survival use hazard rates (HR) – Cox proportional hazards models

Characterization of exposure in cohort studies:

* Regardless of what type of cohort study is conducted, participants are grouped according to their **exposure** and followed over time in order to compare the incidence of symptoms, disease, or death.
* Exposed group: also known as ‘Index Group’
* Unexposed group: also know as ‘Referent Group’.
* When we look at cumulative incidence (risk), we will see that cumulative incidences (risks) can be divided by each other (cumulative incidence in exposed / cumulative incidence in unexposed). This equals the relative risk (RR) which can be calculated in cohort studies. The numerator is the exposed group (index group), while the denominator is the unexposed group (referent or comparison group). Cumulative Incidence is simply a term for RISK. Bottom line, when you hear the term ‘risk’, it refers to the measure of cumulative incidence (CI).
* Thus, it is often the case (but not always … you certainly can put the unexposed group as your index group, provided you make that clear in calculations and explanations) that the unexposed group is the referent group and will often be found in the denominator of a ratio or to the right end of a subtraction. See below.
* CI index group = relative risk (RR) CI referent group
* Or, calculate the risk difference (RD):
* CI index group – CI referent group = risk difference (RD)

Cohort study design: where the 2x2 table comes from

* exposed -> disease or no disease
* unexposed -> disease or no disease

Timing of cohort studies:

* Cohort studies can either be prospective or retrospective.

1. Prospective: forward in time (traditional, focus on this one)

* participants grouped on basis of past or current exposure and are followed into the future in order to observe the outcome(s) of interest
* Prospective studies can obtain more detailed information on exposures and other key variables because the investigators are in control of the data collection process and can gather data directly from the study participants. They are also less vulnerable to bias because the outcomes have not occurred when the cohort is assembled and the exposures are assessed

1. Retrospective: back in time (still cohort but going back in time to do it, use records)

* both the exposures and outcome(s) have already occurred by the time the study begins. It studies prior outcomes only.
* Retrospective cohort studies are more efficient, as they take less time and money than prospective cohort studies for investigating diseases that take a long time to develop and diagnose. This is because the disease and exposure has already occurred. However, these studies often require old medical records or employment records which may lack the information needed or simply be hard to find.
* The key to remember is that a cohort is followed **forward** in time, regardless if it is a prospective or retrospective study. In a prospective study, the cohort you are studying has not had the outcome of interest yet. However, in a retrospective study, you define your cohort at an earlier point in time, before the outcome occurred. Thus, you are still able to follow the cohort forward in time and see who developed and who did not develop the outcome of interest, as well as observe any losses-to-follow-up.

Strengths of cohort studies:

* Can directly measure disease incidence (risk).
* Maintain temporal sequence.
* Can observe exposure and changes in exposure over time (recall is not a problem).
* Do not need to select non-diseased comparison group (controls).
* Can examine several different outcomes.
* Allow subset analyses and nested studies.
* Allow study of rare exposures.
* Can be hard to recall in retrospective manner

Weaknesses of cohort studies

* Inefficient for rare outcomes
* Expensive and time-consuming
* Inefficient for diseases with long induction and latent periods (disease that take a LONG time to develop – e.g., chronic diseases)
* Attrition (loss-to-follow-up) of subjects
* Technology may change (didn’t measure right factors at outset – they may have been diseased but didn’t detect it)
* Loss-to-follow up can reduce the power of your study to detect associations that are truly present because of the smaller sample size available for analysis and it can bias the study results.
* If waiting long time, inefficient. Can run out of funds. Longest are 5-7y and cost millions. Randomized no more than a year. Anything prospective more money

Cumulative incidence (risk)

* Incidence: number or proportion of **new** cases/events
* When you hear the term ‘**risk**’, this is the measure of association being referred to.
* These terms are synonymous with incidence and appear in numerous epidemiology textbooks and research articles, so be familiar with them and know they have the same meaning.
* Incidence
* Cumulative Incidence
* Incidence Proportion
* Risk
* =(#of new cases of dz)/(#in population at risk) over a specified period of time
* The proportion of a fixed population at risk of disease that becomes diseased over a specified period of time.
* The numerator is also included in the denominator, thus making it a proportion.
* Average risk of getting a disease over a specified period of time.
* Proportion: The numerator is included in the denominator. In other words, since the new cases were at one time at risk, they are in both the numerator and denominator.
* Fixed population: has few or no losses to follow-up.
* The critical thing is that the people in the denominator are **at risk of the disease or event**. For example, if you are studying uterine cancer, men would **not** be included in your population at risk. Also, women who have had a hysterectomy or those who already have uterine cancer would **not** be included in the population at risk.
* Must specify time-period
* People in denominator (population at risk) must be **free of disease** of interest at start of observation
* Advantages of Cumulative Incidence:
  + Gives average risk of population
  + Provides an estimate of individual risk

2x2 table how to calculate risk: eg 1 year cohort study of oral contraceptive use and coronary heart dz

|  |  |  |  |
| --- | --- | --- | --- |
|  | CHD | No CHD | TOTAL |
| Oral Contraceptive  (OC) Use | 30  **a** | 30  **b** | 60  **a+b** |
| No Oral Contraceptive Use | 20  **c** | 70  **d** | 90  **c+d** |
| TOTAL | 50 | 100 | 150 |  |

* CI1 (risk of CHD in OC users) = 30 new CHD cases / 60 at risk = 0.5 or 50% (ie risk in exposed)
* CI0 (risk of CHD in non-users) = 20 new CHD cases / 90 at risk = 0.22 or 22% (ie unexposed)

Relative risk:

* Simply stated, relative risk is one risk divided by another risk.
* ‘Relative’ implies **division**.
* Remember that cumulative incidence is ‘risk’.
* Cohort studies divide people by exposure at the beginning, so relative risk is usually the risk in the exposed divided by the risk in the unexposed.
* divide the cumulative incidence (risk) in the exposed by the cumulative incidence (risk) in the unexposed
* RR = CI1 / CI0
* CI1 = risk in exposed

= # new exposed cases / # exposed at risk

* CI0 = risk in unexposed

= # new unexposed cases / # unexposed at risk

* For the previous OC eg:
* Cumulative Incidence (Risk) in exposed (CI1) = a / a+b
* Cumulative Incidence (Risk) in unexposed (CI0) = c / c+d
* Relative Risk (RR) = risk in exposed / risk in unexposed = (a/(a+b)) / (c/(c+d))
* Relative Risk (RR) = Risk in Exposed / Risk in Unexposed = 0.50 / 0.22 = **2.27**
* What is the interpretation of this relative risk? **The risk of developing CHD among OC users is 2.27 times the risk of developing CHD among non-users**
* 50% of OC users (30/60) developed coronary heart disease.
* 22% of non-users (20/90) developed coronary heart disease.
* Overall, 33% of women (50/150) developed coronary heart disease
* If we examine the proportions (or percentages) of OC users among the women with CHD and either the women without CHD or the total group, we see that the percentage of OC users is greater among the women who developed CHD than it is among the women who did not, and therefore among the total group. **So the distribution of OC use (in this case, the distribution is simply the percentage who used OC and its inverse, the percentage who did not) differs according to the value of another factor (CHD).**
* We could also look across the rows here. Notice that 30 of the 100 women in the ‘No CHD’ group use oral contraceptives (30%). Since this is much different than the 60% of CHD women using oral contraceptives, we can see a definite association between CHD and OC use.
* Knowing the value of CHD (i.e., CHD present or CHD absent) gives us information about the distribution of OC use. If we were betting that a randomly selected woman was an OC user, we would want different odds if we were told that she had CHD than if she did not have CHD.
* Association is reciprocal, so we can also describe the association in this table by talking about how the percentage of women who developed CHD differs by OC history. Note, however, that describing an association in a table is not the same as describing incidences or risks. For example, it is most unlikely that half of the women who used OC (30/60) developed CHD. Most likely these data (had I not made them up) would come from a study that began with CHD cases and a sample of non-cases, which is why such a large proportion of the women had developed CHD.
* **This is a positive association** because a greater proportion (50%) of the exposed (OC users) developed CHD than the unexposed (22% of the non-users developed CHD).‘Positive association’ is a simple way of saying that if you are exposed, you are more likely to develop the disease (outcome).

|  |  |  |  |
| --- | --- | --- | --- |
|  | CHD | No CHD | TOTAL |
| Oral Contraceptive (OC) Use | 15  **a** | 30  **b** | 45  **a+b** |
| No Oral Contraceptive Use | 35  **c** | 70  **d** | 105  **c+d** |
| TOTAL | 50 | 100 | 150 |

**33%** of OC users (15/45) developed coronary heart disease.

**33%** of non-users (35/105) developed coronary heart disease.

Overall, **33%** of women (50/150) developed coronary heart disease

- If we look at the percentages of OC users in each of the CHD groups, we see that it is the **same** for the women with CHD AND the women without CHD, and therefore for the total group. Knowing whether we are talking about the women with CHD or the women without CHD gives us no information about the proportion of OC use. Because the same percentage (proportion) of women with CHD and without CHD use oral contraceptives (30% of the CHD cases AND 30% of the non-CHD people use oral contraceptives).

- **No association implies a relative risk (RR) = 1.** The proportion of OC users developed CHD (33%) is equal to the proportion of non-users who developed CHD (33%). Thus, whether or not they use oral contraceptives gives no clue as to who developed CHD.

Negative relative risk association:

|  |  |  |  |
| --- | --- | --- | --- |
|  | CHD | No CHD | TOTAL |
| Oral Contraceptive  (OC) Use | 10  **a** | 50  **b** | 60  **a+b** |
| No Oral Contraceptive Use | 45  **c** | 45  **d** | 90  **c+d** |
| TOTAL | 55 | 95 | 150 |

16% of OC users (10/60) developed coronary heart disease.

50% of non-users (45/90) developed coronary heart disease.

Overall, 37% of women (55/150) developed coronary heart disease

- If we examine the proportions (or percentages) of OC users among the women with CHD and either the women without CHD or the total group, we see that the percentage of OC users is greater among the women who developed CHD than it is among the women who did not, and therefore among the total group. **So the distribution of OC use (in this case, the distribution is simply the percentage who used OC and its inverse, the percentage who did not) differs according to the value of another factor (CHD).**

We could also look across the columns here. Notice that 50 of the 95 women in the ‘No CHD’ group use oral contraceptives (53%). Since this is much different than the 18% of CHD women (10 of 55) using oral contraceptives, we can see a definite NEGATIVE association between CHD and OC use.

- Knowing the value of CHD (i.e., CHD present or CHD absent) gives us information about the distribution of OC use. If we were betting that a randomly selected woman was an OC user, we would want different odds if we were told that she had CHD than if she did not have CHD.

- Higher prop of non users dev dz here, so RR less than 1

Incidence rate (incidence density)

* Incidence rate (IR): how fast or how slow new cases/events are occurring
* Remember basic rates have something per unit time (e.g., miles per hour (MPH), beats per minute (BPM).
* The following terms have the same definition.
* Incidence Rate (IR)
* Incidence Density (ID)
* A true incidence rate has the unit **TIME.** That is what a true incidence rate is measuring, how fast or slow is a disease (event) occurring in a population
* = (#new cases of disease) / (Σ person-time spent at risk in candidate population)
* Numerator for incidence rate is identical to that of cumulative incidence
* The denominator is expressed as the sum of person-time at risk for becoming a case
* Person-time at risk: the amount of disease-free time contributed, summed over every person in the population
* Σ indicates summation (sum of)
* Think of person-time like the years of experience on a basketball coaching staff. You may hear that the University of Toledo coaching staff has 50 years of coaching experience. What does that mean? It means that if you added up all the years that each coach has spent coaching, it would total 50 years. That is the EXACT way person-time is calculated
* In person-time, each unit of time (could be day, month, year, etc.) that an at-risk person spends in the study without developing the outcome is added into the final total of the denominator. For example, let’s say you do a 1-year study on the development of a first myocardial infarction (MI) among 1,000 men. If no men have a MI during that 1-year study period, then each man would contribute 1 person-year of MI-free time to the denominator. Thus, the total for the denominator would be 1,000 person-years.
* If however, 50 men had a MI after exactly 6 months (pretend they all had a MI on June 30th), then these 50 men would only contribute 0.5 years each to the denominator. Why only 0.5 years? Because once a man has his first MI, he is no longer at risk of having another first MI. Get it? The remaining 950 men who went the entire period disease-free would each contribute 1 person-year. Thus, your total person-time for the denominator would be (50\*0.5) + 950 = 975 person-years.
* If asked to calc IR they’re asking for the RR
* A person who never experiences the disease contributes 100% of her/his time to the denominator.
* The time spent with disease is **not included** in the denominator (because a diseased person is not at risk of getting disease during that time).
* Thus, a person no longer contributes person time when she/he gets the disease, the study ends, or is lost to follow-up.
* In practice, including most reports, the incidence rate is typically used to describe the number of new cases that develop in a year in a specific population

Case control studies:

Objectives:

1. Know the general characteristics
2. How do they compare to cohort studies?
3. How do we select controls?
4. Brief review of the odds ratio and confounding
5. Know the advantages and disadvantages

Case control studies

* Identify individuals with the disease (cases) and individuals without disease (controls).
* Next, investigators obtain and compare exposure histories of the cases and controls.
* If there is a positive association between exposure and disease, the prevalence of exposure should be higher among cases than among controls.
* In a Case-Control study, a sample of individuals are selected based on ***case* status**, and then the prevalence of exposure is determined for each group
* For a case-control study we obtain all or a sample of these cases and a sample of controls, and then we compare the exposure distribution in the cases to that in the controls. Restrictive criteria leads to fewer classification errors. For example, if only chest pains were used to define cases of myocardial infarction, most heart attack cases would be included (but people with “silent” ones would be missed= people with other conditions that produce chest pain would be missed). The goal of case identification is to identify as many **TRUE** cases (prevalent and incident cases) of disease quickly and inexpensively. Thus, the ability to achieve this goal depends on time and money (resources available).
* It's generally most desirable to identify incident cases as close to the event (disease onset) occurs. Incident cases are preferable over prevalent cases for causal research because investigators are usually interested in factors that lead to developing the disease rather than factors that affect the disease duration.

When to use case-control design?

* Exposure data are difficult or expensive to obtain.
* The disease is rare.
* The disease has a long latent or induction period (e.g., most chronic diseases).
* Little is known about the disease.
* The underlying population is dynamic.
  + Collecting exposure data on fewer people saves time and money.
  + Fink was taught a disease is considered rare if the incidence is less than 10%. This is a judgment call.
  + **For example**, cancers involving solid tumors may have a latency period (time between point of pre-clinical detection to disease diagnosis) of 15 years or more. It is hard to follow people for that much time and is very expensive.

Selection of controls

* Controls should estimate the distribution of exposure in the source population
* Controls should be sampled independently of exposure status
* In other words, controls should be as similar as possible to the cases, without having the disease/outcome.
* Types
  + Hospital / clinic based controls
  + Population controls
  + Neighborhood controls
  + Random digit dialing
* What does it mean that controls should estimate the distribution of exposure in the source population? That the controls should be selected independently of their exposure status. The reason for this is that we can use the control groups in place of the denominator information in incidence rates (person-time) to determine the ratio of the incidence rate in exposed people relative to the incidence rate in the unexposed people (Incidence Rate Ratio or IRR). If you select controls at the end of the study period, they likely will have more **unexposed** people if the exposure is a risk factor for disease. Recall, IRs and IRRs are common calculations in cohort studies. Thus, we can estimate incidence rate ratios (IRR) from an odds ratio if our controls estimate the distribution of exposure in the source population.
* These are some general rules for control selection.

1. Controls should be selected from the same population (the source population) that gives rise to the cases.
2. Controls should be selected independently of their exposure status. That is, the controls should be representative of the source population with respect to exposure.
3. The time during which a subject is eligible to be a control should be the time in which that individual is also eligible to become a case, if the disease should occur. Thus, an individual in whom the disease has already developed or who has died is no longer eligible to be selected as a control.

* A key feature of the selection of controls is their continuous eligibility to become cases if they should develop the disease.
* **Hospital / clinic-based Controls.**
  + Easy to capture, but reference population (from which cases arose) can be difficult to define.
  + Which diseases or events are suitable for control group?
    - * + Illness should be unrelated to exposure.
        + Illness in the controls should have same referral pattern as cases.
* **Population Controls**
  + Identified from school rosters, insurance company lists, dept of motor vehicles, Medicare beneficiaries.
  + Neighborhood controls; door-to-door, random-digit dialing.
* Investigators are usually assured that population based controls come from the same base as the cases.
* Challenges of population based controls
  + Expensive
  + Time consuming
  + Recall of past exposures limited
  + Cooperation of population based controls may be limited.
* A key feature of the selection of controls is their continuous eligibility to become cases if they should develop the disease. This is preferred for studying etiology and acquiring incident cases.
* **For example**, if the study period spans three years. A person free of disease in year 1 is a potential control and may go on to develop the disease in year 3, thus becoming a case. The control group in case-control studies is intended to provide estimates of the relative size of the denominators of the incidence rates for the compared groups. Therefore, each case in a case-control study should have been eligible to be a control before the time of disease onset. Each control should be eligible to become a case as of the time of selection as a control. An individual selected as a control who later develops the disease and is selected as a case should be included as **BOTH** a control and a case. Remember cohort studies? Each person who develops the disease would be in the numerator and the denominator. J This is a similarity between the two study types.

Odds:

* Odds is simply the number of ways an event can occur, divided by the number of ways an event cannot occur.
* For example, if you have 5 marbles in a jar (1 blue and 4 red), what is the odds of selecting the blue marble on your first try?
* Odds = 1 blue marble / 4 red marbles or 1:4. This is because for every blue marble you could select, there are 4 red marbles.
* The probability of choosing the blue marble is 1 / 5 = 20%. Notice the blue marble is in both the numerator and denominator here.
* More examples:
  + Horse racing lists odds of winning, 3:1, 5:2, 10:1, 20:1 75:1.
  + In a standard deck of 52 cards, the odds of picking a diamond is 13:39 or 1:3.
  + For every diamond, there are 3 cards that are not diamonds.

Odds and Probability

* Odds = Probability / (1 – Probability) = P / (1-P)
* Example: If there are 20 cases (15 that are exposed to cigarette smoke and 5 that are unexposed), what is the probability AND what is the odds of being **unexposed**?
* Odds = 0.25 / (1 – 0.25) = 0.25 / 0.75 = 0.33 or 1:3. In other words, there will be 1 unexposed person for every 3 exposed persons
* Odds CANNOT be expressed as a percentage because it is **NOT** a proportion.
* So what does an odds ratio (OR) mean?
* Let’s look at the odds of an event. This is the ratio of the number of ways an event can occur divided by the number of ways an event CANNOT occur. Suppose we bet on a horse, Smarty Jones, to win a race. Smarty Jones has a 60% probability (risk) of winning the race (P=0.60). Thus, he has a 40% probability of losing the race (1 – P = 0.40). What are the **odds** that Smarty Jones wins the race.
* Odds = P / (1 – P) = 0.60 / (1 – 0.60) = 1.5 : 1 = 1.5
* P is the probability or risk of the outcome, like the risk of developing the disease.
* O is the odds of the outcome, like the odds of developing disease.
* Notice the distinction between the **probability (risk)** of winning = 60%. The **odds** of winning are 1.5 : 1 or 1.5.
* Not same prob can be shown as %

Odds ratio:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Case | Control | TOTAL |
| Exposed | **a** | **b** | a + b |
| Unexposed | **c** | **d** | c + d |
| TOTAL | a + c | b + d | N |

Odds of being exposed in cases/odds being exposed in controls = (a/c)/(b/d)

Mathematically same as cross product

* So what does an odds ratio (OR) mean?
* The **exposure odds ratio** equals the odds of being exposed among the cases divided by the odds of being exposed among the controls. For example if the OR = 1.50 in a study of HRT use and breast cancer, the odds of being exposed to HRT among breast cancer cases are 1.50 times the odds of being exposed to HRT among controls. Exposure odds ratio = (a/c) / (b/d)
* The **disease odds ratio** equals the odds of developing the disease among the exposed divided by the odds of developing the disease among the unexposed. For example, for the OR = 1.50, the odds of developing breast cancer in those taking HRT are 1.50 times those for those not taking HRT.
* Case control study look at odds and odds ratio. Do NOT calc rel risk here
* OR = (a/c) / b/d
* If the OR=1, there is no association between exposure and case/control status.
* If the OR>1, Positive association.
* If the OR<1, Negative association.

Strengths of case-control studies:

* Less cost
* Less time
* Fewer subjects
* Suitable for rare diseases
* Suitable for diseases with long latent periods
* May simultaneously investigate several exposures

Weaknesses of case-control studies

* Limitations in recall may lead to misclassification
* Problems in control selection
* Only estimates RR and IRR, and only under certain conditions
* Inefficient for rare exposures